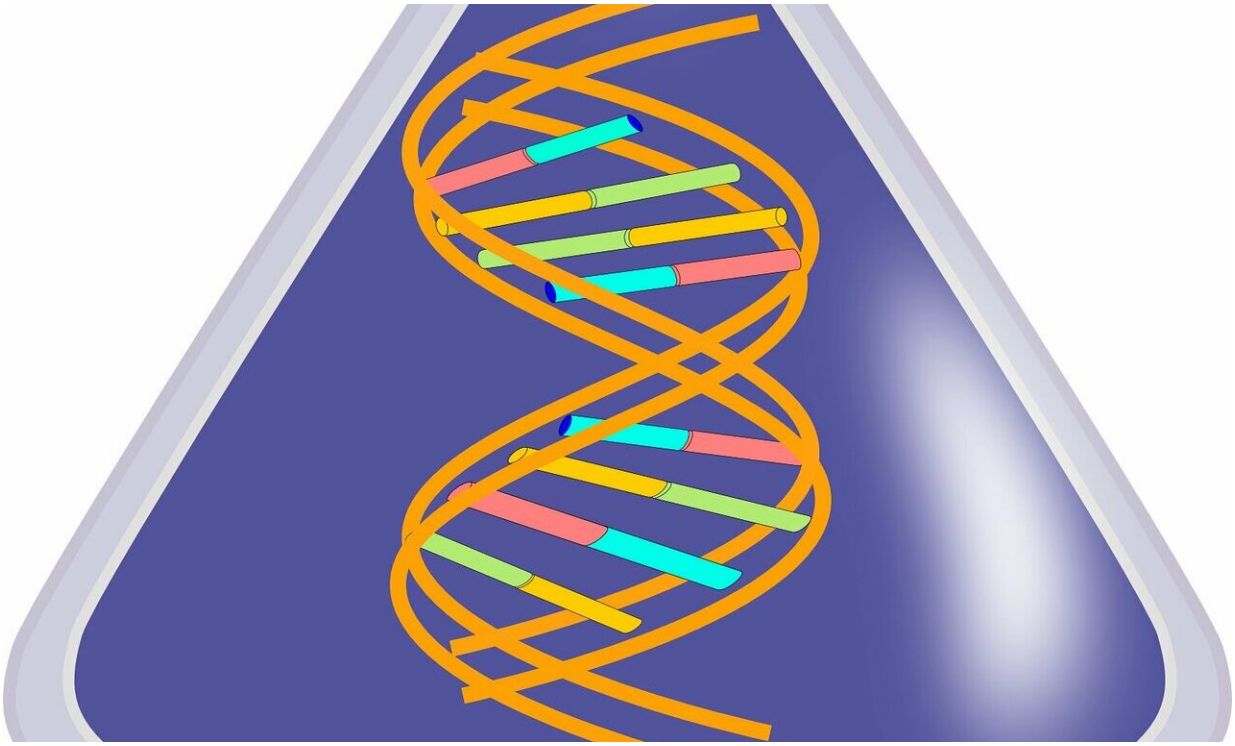
Widespread misinterpretation of gene expression data

November 12 2019



Credit: CC0 Public Domain

Reproducibility is a major challenge in experimental biology, and with the increasing complexity of data generated by genomic-scale techniques this concern is immensely amplified. RNA-seq, one of the most widely used methods in modern molecular biology, allows in a single test the simultaneous measurement of the expression level of all the genes in a

given sample. New research publishing November 12 in the open-access journal *PLOS Biology* by Shir Mandelbaum, Zohar Manber, Orna Elroy-Stein, and Ran Elkon from Tel Aviv University, identifies a frequent technical bias in data generated by RNA-seq technology, which recurrently leads to false results.

Analysing dozens of publicly available RNA-seq datasets, which profiled the cellular responses to numerous different stresses, Mandelbaum and colleagues noticed that sets of particularly short or long genes repeatedly showed changes in expression level (as shown by the apparent number of RNA transcripts from a given gene).

Puzzled by this recurring pattern, the authors then asked whether it reflects some universal biological response common to many different triggers or it rather stems from some experimental artefact. To tackle this question, they compared replicate samples from the same biological condition. Differences in gene expression between replicates can reflect technical effects that are not related to the experiment's biological factor of interest. Unexpectedly, the same pattern of particularly short or long genes showing changes in expression level was observed in these comparisons between replicates, demonstrating that this pattern is the result of a technical bias that seemed to be coupled with gene length.

A main goal of RNA-seq experiments is to characterize biological processes that are activated or repressed in response to the conditions of interest. Notably, specific biological processes are executed by products of particularly short and long genes. For example, many of the short genes encode proteins that constitute the ribosome, the cell's protein-making machinery. Conversely, many of the long genes encode proteins that constitute the extra-cellular matrix (ECM), the network of macromolecules that provide cells with an external structural support.

Mandelbaum and colleagues were able to show how, in many RNA-seq

datasets, the length bias they detected, combined with some flaws in the statistical analysis, can lead to the false identification of specific biological functions (including ribosome and ECM-related functions) as cellular responses to the conditions tested. Importantly, the study also shows how this bias can be removed from the data, thus filtering out false calls while preserving the biologically genuine ones.

Recent years have witnessed a growing alarm about false results in biological research, sometimes referred to as the reproducibility crisis. This study emphasizes the importance of proper statistical handling of data to lessen the number of misleading findings.

More information: Mandelboum S, Manber Z, Elroy-Stein O, Elkon R (2019) Recurrent functional misinterpretation of RNA-seq data caused by sample-specific gene length bias. *PLoS Biol* 17(11): e3000481. doi.org/10.1371/journal.pbio.3000481